

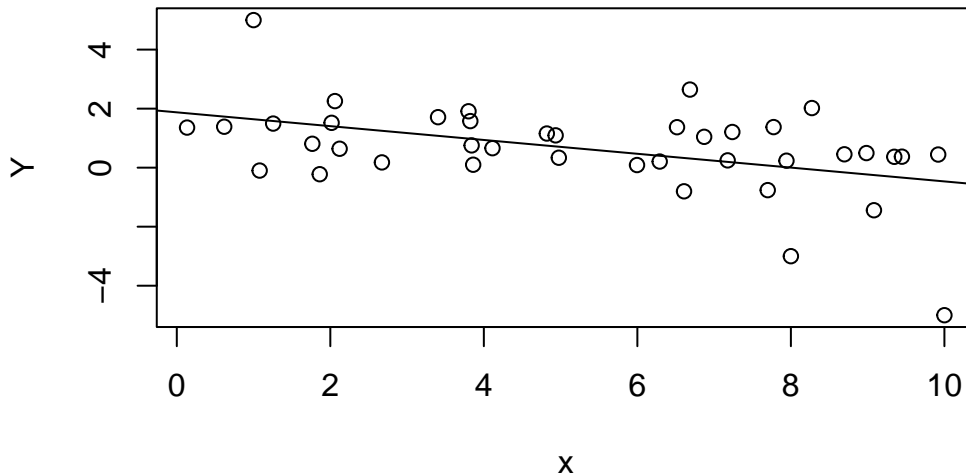
# STAT 516 sp 2024 exam 01

75 minutes, no calculators or notes allowed

## 1. Simple linear regression

Below is a scatterplot of  $n = 40$  data points  $(x_1, Y_1), \dots, (x_{40}, Y_{40})$  with the least squares line overlaid.

```
plot(Y~x)
abline(lm(Y~x))
```



```
summary(lm(Y~x))
```

Call:

```
lm(formula = Y ~ x)
```

Residuals:

```

      Min       1Q   Median       3Q      Max
-4.5336 -0.6735  0.0845  0.7369  3.3597

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.87442     0.45380   4.131 0.000191 ***
x            -0.23409     0.07469  -3.134 0.003315 **

```

---

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 1.389 on 38 degrees of freedom

Multiple R-squared: 0.2054, Adjusted R-squared: 0.1845

F-statistic: 9.824 on 1 and 38 DF, p-value: 0.003315

Some of the following questions refer to the above R output; some are general questions that you can answer without referring to the R output.

(a) What do we call the quantity  $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$  and what does it represent?

(b) What do we call the quantity  $\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$  and what does it represent?

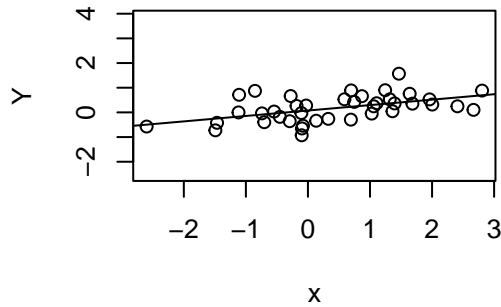
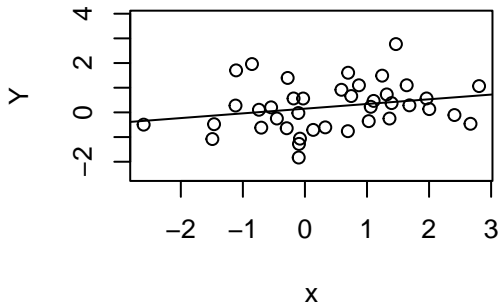
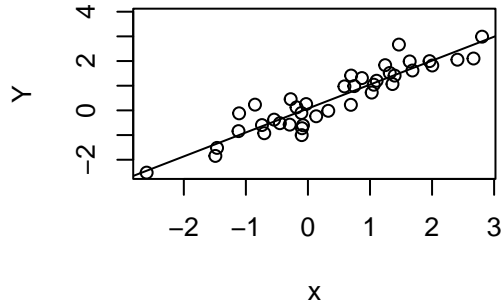
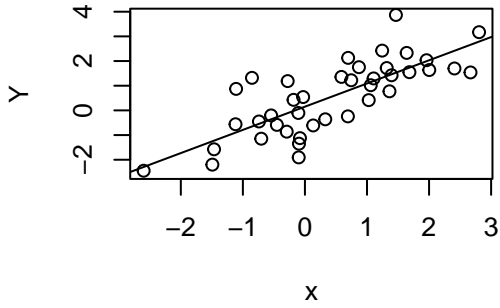
(c) Give the value shown in the R output for  $\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}$ . Interpret the value.

(d) Obtain the value of  $\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}$  from the R output. What does it estimate?

- (e) Give the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  from the R output. Give an interpretation of  $\hat{\beta}_1$ .
- (f) Confidence intervals for  $\beta_0 + \beta_1 x_{\text{new}}$  as well as prediction intervals for  $Y_{\text{new}}$  at new values of the predictor  $x_{\text{new}} = 5$  and  $x_{\text{new}} = 9$  are given below. For each interval, indicate whether it is a CI or a PI and indicate to which value of  $x_{\text{new}}$  it corresponds.
- i. (-3.13, 2.67)
  - ii. (0.26, 1.15)
  - iii. (-0.94, 0.48)
  - iv. (-2.14, 3.55)
- (g) Circle a data point on the scatterplot which would have a large value of Cook's D. Explain your choice of data point.
- (h) Circle a data point on the scatterplot which would have a small value of Cook's D. Explain your choice of data point.
- (i) There is a p-value which appears twice in the R output. Explain why the same p-value appears twice.

(j) Scatterplots of four different data sets are shown below. Indicate for which data set the value of  $F_{\text{stat}} = \frac{MS_{\text{Reg}}}{MS_{\text{Error}}}$  would be

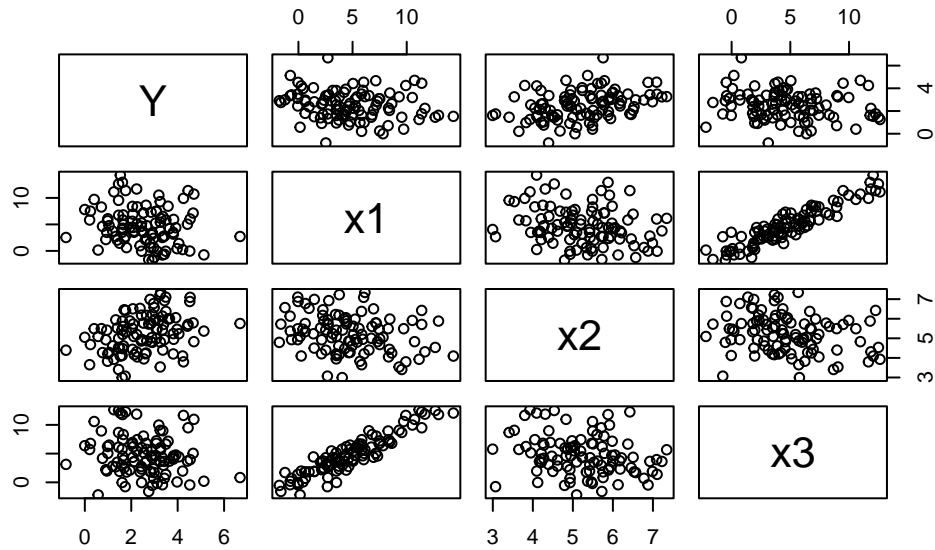
- a. the greatest.
- b. the smallest.



## 2. Multiple linear regression

The plot below shows scatterplots between all pairs of variables in a data set. Following that is some regression output.

```
plot(data)
```



```
lm1 <- lm(Y ~ x2, data = data)
summary(lm1)
```

Call:

```
lm(formula = Y ~ x2, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.9448	-0.7886	0.0424	0.6243	3.9408

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.1536	0.6588	0.233	0.816083
x2	0.4506	0.1237	3.643	0.000433 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.188 on 98 degrees of freedom

Multiple R-squared: 0.1193, Adjusted R-squared: 0.1103

F-statistic: 13.27 on 1 and 98 DF, p-value: 0.0004333

```
lm2 <- lm(Y ~ x1 + x2 + x3, data = data)
summary(lm2)
```

```
Call:
lm(formula = Y ~ x1 + x2 + x3, data = data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.0473 -0.8223 -0.0535  0.6444  3.9421
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.43876    0.74379   0.590  0.55664
x1          -0.05382    0.08834  -0.609  0.54385
x2           0.42276    0.12841   3.292  0.00139 **
x3           0.02437    0.09137   0.267  0.79025
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.194 on 96 degrees of freedom
Multiple R-squared:  0.1277,    Adjusted R-squared:  0.1004
F-statistic: 4.684 on 3 and 96 DF,  p-value: 0.00426
```

Use the above R output to answer the following questions.

(a) For the model with all three predictors, give the value of each entry in the ANOVA table:

Source	Df	SS	MS	F value	p-value
Regression	i.	ii.	iii.	iv.	v.
Error	vi.	vii.	viii.		
Total	ix.	x.			

Since you may not use a calculator, give expressions that could be evaluated in order to obtain the right numbers! Two are already completed in this way as examples.

i.

ii.

iii.

iv.

v.

vi.

vii.  $(1.194)^2 \cdot 96 = SS_{\text{Error}}$

viii.  $(1.194)^2 = MS_{\text{Error}}$

ix.

x.

(b) Which two predictor variables will have the highest variance inflation factors? How can you tell?

(c) For the model with all three predictors, give the null and alternate hypotheses for the overall F-test of significance.

(d) Suppose we wish to test simultaneously the significance of  $x_1$  and  $x_2$ . Write down the relevant null and alternate hypotheses.

(e) Give the value of  $s$  needed to compute the test statistic

$$F_{\text{stat}} = \frac{(SS_{\text{Error}}(\text{Reduced}) - SS_{\text{Error}}(\text{Full}))/s}{SS_{\text{Error}}(\text{Full})/(n - (p + 1))}$$

of the full-reduced model F-test.

- (f) The value of the test statistic  $F_{\text{stat}}$  for the full-reduced model F-test is 0.464. Moreover,  $F_{2,96,0.05} = 3.091$ . What do we conclude about the significance of x1 and x2?

### 3. Inference on the mean of a Normal distribution

Let  $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$  and suppose we wish to test  $H_0: \mu = 1$  versus  $H_1: \mu \neq 1$ . Let

$$T_{\text{stat}} = \frac{\bar{X}_n - 1}{S_n/\sqrt{n}}$$

and suppose we reject  $H_0$  when  $|T_{\text{stat}}| > t_{n-1, \alpha/2}$  for some significance level  $\alpha$ . Answer the following questions about the probability  $P(|T_{\text{stat}}| > t_{n-1, \alpha/2})$ , which is the probability of rejecting  $H_0$ , also called the *power* of the test.

- (a) Suppose  $\mu$  is truly equal to 1. Then give  $P(|T_{\text{stat}}| > t_{n-1, \alpha/2})$

- (b) What happens to  $P(|T_{\text{stat}}| > t_{n-1, \alpha/2})$  as  $\mu$  moves away from 1?

- (c) Suppose  $\mu$  is not equal to 1. What happens to  $P(|T_{\text{stat}}| > t_{n-1, \alpha/2})$  if the sample size is increased?



(d) Suppose  $\mu$  is not equal to 1. What is the effect on  $P(|T_{\text{stat}}| > t_{n-1, \alpha/2})$  of a larger variance  $\sigma^2$ ?

(d) Suppose  $\mu$  is truly equal to 1. What is the effect on  $P(|T_{\text{stat}}| > t_{n-1, \alpha/2})$  of a larger sample size  $n$ ?